Path-Specific Effects

Weinberger, Naftali

Published in:
The British Journal for Philosophy of Science

Publication date:
2019

Document Version
Peer reviewed version

Path-Specific Effects

Naftali Weinberger

Forthcoming in *The British Journal for the Philosophy of Science*
Unformatted Version

Abstract: A cause may influence its effect via multiple paths. Paradigmatically (Hesslow, 1974), taking birth control pills both decreases one's risk of thrombosis by preventing pregnancy and increases it by producing a blood chemical. Building on Pearl (2001), I explicate the notion of a *path-specific effect*. Roughly, a path-specific effect of C on E via path P is the degree to which a change in C would change E were they to be transmitted only via P. Facts about such effects may be gleaned from the structural equations commonly used to represent the causal relationships among variables. I contrast my analysis of the Hesslow case with those given by theorists of probabilistic causality, who mistakenly link it to issues of causal heterogeneity, token-causation and indeterminism. The reason probabilistic theories misdiagnose this case is that they pay inadequate attention to the structural relationships among variables.

0. Introduction

One of the most widely discussed causal scenarios remains one of the least understood. In the paradigm version (Hesslow, 1974), taking birth control pills both decreases one's risk of thrombosis by preventing pregnancy – itself a cause of thrombosis – and increases the risk of thrombosis by producing a certain blood chemical. Here there is an inclination to distinguish between the *net* effect of birth control on thrombosis, and the effects of birth control on thrombosis *via* the specific paths through pregnancy and the blood chemical. Yet it remains unclear both how to define path-specific effects and, as importantly, how such effects relate to the net effect. The question of how to understand multi-path cases was extensively debated by theorists of probabilistic causality, who mistakenly linked it to questions regarding causal heterogeneity, token-causation and indeterminism. Here I build on Pearl's (2001) treatment of these cases to provide a philosophical analysis of path-specific effects. This analysis uses the causal counterfactuals represented by structural equations to capture how changes in pregnancy status are transmitted along different causal paths. The analysis makes precise how path-specific effects contribute to net effects, and illustrates how structural equations enable one to address questions that eluded probabilistic accounts.

The paper is organized as follows. Section 1 introduces probabilistic and structural equations accounts of causality, reviews standard challenges to

probabilistic accounts, and compares the accounts' treatment of confounding. Section 2 delves into a debate between Ellery Eells and Nancy Cartwright to illustrate how theorists of probabilistic causality linked their analysis of multi-path cases to issues of type versus token causality, causal heterogeneity and indeterminism. Section 3 defines path-specific effects and provides an alternative analysis of these cases. Section 4 highlights the differences between the proposed account and probabilistic ones, and clarifies how the former diverges from Hitchcock's (2001b) superficially similar account. Section 5 concludes.

1. Probabilistic Causality and Structural Causal Models

The core idea shared by theories of probabilistic causality is that causes increase the probabilities of their effects (Good (1961), Suppes (1970), Cartwright (1979), Skyrms (1980), Eells and Sober (1983), Eells (1991)). More precisely, causes raise their effects' probabilities relative to certain *background contexts* – i.e. sets of variables one conditions on. On Eells' account, for example, a background context for evaluating the effect of X on Y specifies the values of all variables that are causally relevant to Y that are not effects of X. He maintains that X is a *positive* causal factor for Y in a population if and only if $P(Y|X\&K)$ is greater than $P(Y|\sim X\&K)$ in all background contexts K in that population. It is a *negative* factor if it lowers Y's probability in these contexts, and *neutral* if it does not change the probability of Y. If X raises the probability of Y in some contexts and lowers it in others, it has *mixed* causal relevance. Note that causal roles – whether a cause has positive, negative or mixed relevance for its effect – are population-relative; smoking may be positively relevant to cancer in one population, but not another.

It will help to distinguish among three types of challenges to probability-raising accounts. The first is that of specifying sets of background factors relative to which probability raising indicates causation. Since X may raise the probability of Y not because X causes Y, but rather due to a common cause, one must include any common cause of these variables in the background context. As noted, Eells goes further and says that one must include other causes of Y as well. The second challenge involves cases such as Hesslow's in which X influences Y along multiple paths. Most saliently, the distinct paths might cancel, so that there will be no probabilistic relationship between X and Y despite there being chains of causal influence. Among other issues, these cases highlight the difficulty for probabilistic accounts in guaranteeing the transitivity of causation. Third, there appear to be probability-lowering causes of an effect. In Deborah Rosen's (1978) example, a golfer hits a ball, which on its way towards the hole is kicked by a squirrel. Suppose that the squirrel's kick lowers the probability of the ball going into the hole, but the ball still does. Although the kick lowered the

probability of success, intuitively it caused it. One might insist that the kick does increase the probability of success relative to a finer-grained specification of its properties. But perhaps there is no such specification and the probability of success is irreducibly chancy. The standard probabilistic solution to this problem (Sober, 1984) is to distinguish between "type" and "token" causation, (though the labels "general" and "singular" may be preferable (Hitchcock, 1995)). The squirrel kick is a type-level negative causal factor, but a token-level cause, of the ball's going into the hole.

Regarding the first challenge, probabilistic theorists never fully distinguished between two reasons for including variables in a background context. One is to eliminate confounding. Another is to get unambiguous facts about causal role. Suppose one considers the effect of smoking on cancer relative to a background context that contains all common causes, but omits another cause, *Z*, of cancer. The probabilistic relationship between smoking and cancer could then differ among different members within that context with different values for *Z*. So even if smoking raised the probability of cancer in every (incompletely specified) context, it would not follow that smoking raises the probability of cancer in finer-grained contexts. This is the rationale for conditioning on more than common causes in establishing facts about causal role.

Even so, one need not condition on a complete set of background factors to avoid confounding – a point that probabilistic theorists ought to grant. Yet Eells denies that there is ever a reason to consider incomplete background contexts, as becomes clear from his response to Dupré (1984). Dupré maintains that causal relevance need not be evaluated relative to a complete background context, but merely relative to a "fair sample". Eells justifiably criticizes Dupré for not clarifying what he means by a "fair sample", though some have charitably interpreted Dupré as considering a sample from an experimental population where confounding has been eliminated (Hausman, 2010). In such a population, the effect size – the amount by which the cause raises the probability of its effect – may differ for individuals with different background factors, but the magnitude of the effect in the population is just the average of the individual effect sizes. Eells rejects this proposal, stating that the "average effect is a sorry excuse for a causal concept" (1987, 113). The basis for this rejection is that since the average effect depends on the contingent frequencies of background factors in the population, allowing such effects undermines one's ability to provide an account of causal laws (which do not depend on such contingent factors).

According to Hitchcock (2001a), Eells and Dupré are not disagreeing, but rather are explicating distinct concepts. Eells explicates causal laws, which are evaluated relative to homogenous background contexts. Dupré explicates the causal effects measured in randomized experiments. What Hitchcock's conciliatory approach leaves out is that

neither Eells nor Dupré specify how one can eliminate confounding with an incomplete set of background variables. As noted, Dupré vaguely refers to a "fair sample" and – charitable interpretations aside – he never discusses experiments in the cited article. Given a specification of which incomplete sets of background factors are sufficient for avoiding confounding, one can draw a clear boundary between issues of confounding and those of heterogeneity, as I will presently show. But such a specification is absent from the probabilistic literature. This accounts for the mistrust of average effects even among those who sometimes allow for them. We will see that this mistrust influences probabilistic discussions of the Hesslow case.

Within the causal modeling framework developed by Spirtes, Glymour and Scheines (2000), and Pearl (2009), one can make precise which variables must be held fixed to eliminate confounding and establish causal relevance. These frameworks represent causal hypotheses using directed acyclic graphs, or DAGs, in which the nodes are random variables and the edges are arrows. The arrows represent direct causal relationships.[1] 'Acyclic' indicates that no variable is a cause (either direct or indirect) of itself.

When there is an unmeasured common cause of X and Y, $P(Y|X)$ provides a *biased* measurement of the effect of X on Y.[2] In contrast, if X and Y share a complete and proximate common cause Z, one can learn about the effect of X on Y by conditioning on Z at a particular value. In this simple case, one can get an unbiased measurement of the effect of X on Y by conditioning on the common cause. Common causes are not the only variables that bias an effect measurement. If one conditions on a common *effect* of X and Y, this may also introduce bias. Consider a bathtub in which X is the rate at which water flows into the tub, Y is the size of the drain, and Z is the water's depth. X and Y cause Z, and are probabilistically independent. Yet conditional on the water's depth, the size of the drain is informative about the rate of flow. This reveals how conditioning on a common effect can render its causes probabilistically dependent. Accordingly, any probabilistic dependence between the causes will not reliably indicate the presence of a causal relationship between them.

---

[1] The concept of direct cause (Woodward, 2003 p.55) is distinct from that of the direct effect (section 3). 'Direct cause" is a qualitative concept corresponding to whether there is an arrow in a DAG. 'Direct effect" is a quantitative concept. In what follows, I presuppose knowledge of the correct DAG. While the existence of a direct effect entails a direct cause, the converse does not hold.

[2] $P(Y|X)$ may differ for different values of $X=x$ and $Y=y$ (uppercase italicized letters denote variables, and lowercase italicized letters their values). There are various measures of causal strength that may be inferred from the distribution $P(Y|X)$ (Fitelson and Hitchcock, 2011), though one need not distinguish among these to determine whether one can infer causal information from the probability distribution.

DAGs enable one to generalize these considerations to cases involving any number of variables. A *path* from X to Y is a set of connected arrows between X and Y going in any direction. A *causal path* from X to Y is a set of connected arrows going in the *same* direction (from X to Y). Y is a collider on a path X–Y–Z just in case $X \rightarrow Y \leftarrow Z$. A path is *blocked* iff it contains a collider that one does not condition on[3] or a non-collider that one does condition on. The correlation between X and Y conditional on variable set $\boldsymbol{V}$ provides an unbiased estimate of the effect of X on Y iff $\boldsymbol{V}$ blocks all and only the non-causal paths from X to Y.[4] Crucially, $\boldsymbol{V}$ need not include all causes of Y. Where there are causes of Y that are not linked to X by any unblocked path, these do not induce bias.

Putting this all into language that would be more familiar to Eells, X is causally relevant to Y iff X changes the probability of Y relative to at least one background, $\boldsymbol{V}$, which includes variables blocking all and only the non-causal paths between X and Y. This is not the standard formulation of causal effects within the more recent frameworks. A more common formulation is that X causes Y if it is possible to change the value of Y via an ideal intervention on X. More precisely, let $P(Y|\text{do}(X))$ be the probability of Y given an intervention on X. X is causally relevant to Y if there are at least two values of X, x and x' such that $P(Y|\text{do}(X=x)) \neq P(Y|\text{do}(X=x'))$. An ideal intervention on X sets it to a particular value such that its value depends only on the intervention and not any of X's other causes (see Pearl, 2009; Woodward, 2003). This explication of causes using ideal interventions elucidates how one can establish causal claims experimentally. Yet talk of interventions is dispensable: when one's background $\boldsymbol{V}$ is sufficient for getting an unbiased estimate, $P(Y|\text{do}(X), \boldsymbol{V}) = P(Y|X, \boldsymbol{V})$ for all values of X, Y and $\boldsymbol{V}$.[5] Accordingly, X raises the probability of Y relative to $\boldsymbol{V}$ iff intervening on X would increase the probability of Y.

While I have emphasized how graphical models help establish causal relevance, this is not their only – or even their primary – use. Graphical models allow one to represent the functional dependence of effects on their causes using *structural equations*. We may refer to a variable's direct causes in a

---

[3] Conditioning on downstream effects of a collider also unblocks a path containing the collider.

[4] Note that even when $P(Y|X, \boldsymbol{V})$ does not provide an unbiased estimate, the effect of X on Y may be identifiable via an adjustment procedure (Pearl, 2009 p. 79-83).

[5] Pearl (2009, p. 85-6) provides a sound and complete set of rules for when it is possible to replace a do-expression with a non-do-expression. Rule 2 entails that $P(Y|\text{do}(X), \boldsymbol{V}) = P(Y|X, \boldsymbol{V})$ when every non-causal path between X and Y is blocked by $\boldsymbol{V}$ (in applying the rule, one considers a graph in which all arrows out of X are deleted so that there are no causal paths from X to Y). Spirtes, Glymour, and Scheines (2000, pp. 164, 313) give a theorem that is equivalent to Pearl's rules.

graph as its *parents*. For each variable in a graph, there is a structural equation representing that variable as a function of its parents and (typically) an error term. A variable's error term represents its causes that are not included in the model, and which vary in the probability distribution. The form of the functional relationship between a variable and its parents need not be assumed a priori. Given a graph and a probability distribution, this relationship is *identifiable* iff it is uniquely determined from the graph and the probability distribution without intervening. The conditions under which a cause is identifiable are identical to those under which one can get an unbiased effect measurement. The presence of error terms underscores the fact that not *all* causes must be included in a model to identify causal effects. But one must not omit *common* causes. The presence of unmeasured common causes may render some causal quantities non-identifiable. To simplify things, in the following I consider only DAGs over variable sets not omitting any common causes of variables in the set.

DAGs allow for a clean distinction between quantitative and qualitative causal facts. Given a DAG in which $X \rightarrow Y$, the quantitative relationship between X and Y is identified, where possible, based on the probability distribution. It is the probability distribution rather than the qualitative causal relationships in the DAG that determine quantitative facts about causal role. Relatedly, DAGs do not distinguish between homogenous and heterogeneous causal relationships. Suppose that $X \rightarrow Y$ and there is a cause Z of Y that is omitted from the DAG. The relationship between X and Y is homogenous with respect to Z if it is the same for all values of Z and heterogeneous with respect to Z otherwise. When Z is omitted from the graph, any influence that it has on the probabilistic relationships between X and Y is nevertheless reflected in the probability distribution that does not include Z. The graph itself does not specify whether the causal relationship between X and Y is homogenous with respect to Z (or other causes of Y).

One might suppose that heterogeneity is problematic in one version of the Hesslow case. Specifically, the paths by which birth control causes thrombosis might cancel such that birth control is uncorrelated with thrombosis in a population. The paths might nevertheless not cancel in subpopulations with different background factors. It might seem like the absence of a correlation between X and Y in the population would lead one to misidentify the effect of X on Y. Yet, given a DAG in which $X \rightarrow Y$, one would draw the *correct* conclusion that X has no net effect on Y.[6] One might wonder how we would *learn* that X causes Y. But the present point is that

[6] It is non-obvious whether there should be arrow from X to Y when X has no net effect on Y in the distribution, but does influence Y relative other distributions. I write as if we should include the arrow, although nothing here depends on it. If we omit the arrow, there is no causal quantity to identify and no false conclusion to draw about the effect of X on Y.

given a correct DAG, all other facts about the quantitative relationship between the variables are identified from the probability distribution. Heterogeneity raises no barriers to doing so.

Hesslow's case raises a more subtle issue. Can probabilistic theories account for the intuition that birth control does not merely have a net effect on thrombosis, but also influences it via particular paths? And should they? Probabilistic theorists extensively debated these issues. They tried to resolve them using the same tools that they used to address the three challenges enumerated above. As we will see, these were not the right tools for the job.

2. Cartwright and Eells Debate Birth Control

Probabilistic causal theorists were greatly interested in multi-path cases (Hesslow, 1976; Dupré, 1984; Otte, 1985; Cartwright, 1988). Unlike the account I will present, these philosophers did not attempt to quantify the influence of birth control on thrombosis via a path.[7] They nevertheless tried to make sense of the idea that birth control can have a positive effect on thrombosis along one path, and a negative one along another. In chapter 4 of *Probabilistic Causality*, Eells rejects various proposals for unpacking this idea. He grants that birth control is causally relevant to pregnancy and to the blood chemical and that these are relevant to thrombosis. He further grants that these distinct causal relationships can differ in strength across populations and that this can make a difference in the net effect across populations. Yet he denies that there is any population in which birth control is type-level causally relevant to thrombosis in two ways.

Eells resistance was justified, given that no one had advanced a compelling account of the allegedly "dual" nature of this effect. Yet the flaws in these accounts do not show that we cannot make sense of path-specific effects. Rather, they highlight the limitations of probabilistic approaches. The problem with these approaches is that they explicate the dual nature of the effect in terms of there being different populations in which birth control influences thrombosis differently. An oversimplified but illustrative version of such a view would be that in each homogenous population of women, birth control influences thrombosis via one path rather than the other. On this picture, the relationship between the net effect and path-specific effects is akin to that between an average effect and the homogenous effects in the populations that get averaged. Here I focus on Cartwright's analysis of the Hesslow case, which she developed as part of an extended debate with Eells (Eells and Sober 1983; Eells, 1988; Cartwright, 1988a; Cartwright 1988b, Cartwright 1989, Eells, 1991). Her view is more sophisticated than that just suggested. Yet it resembles it in viewing the net effect as an average effect.

[7] Though see Cartwright (1988, pp. 91, 92-7).

Correspondingly, Eells rejects path-specific effects by denying that the net effect is an average effect.

The technical point of contention in Eells' and Cartwright's debate concerns when one should condition on intermediate variables, or *mediators*, in evaluating the effect of birth control on thrombosis. Eells maintains that one should never condition on mediators, although one should condition on all causes of mediators that are not themselves mediators.[8] Eells and Sober (1983, p. 40) provide an example illustrating why one should not condition on mediators. Suppose you call me, which causes the phone to ring, leading me to pick it up. There is no caller ID, so whether I pick up is not influenced by whether it is you calling. We should not say that your calling me is relevant to my picking up only if it changes the probability of my picking up *conditional* on the phone's ringing. Since your calling me influences my picking up *via* the phone's ringing, we should not hold the phone's ringing fixed in evaluating the effect. While this example involves only a single path, the reasoning generalizes to multi-path cases.

In contrast, Cartwright holds that one sometimes should hold fixed the values of mediators. In Hesslow's case we should hold fixed the pregnancy variable for those women for whom birth control was not a singular cause of their not getting pregnant. Cartwright's reasoning for why we must appeal to singular causation goes as follows.[9] Consider the influence of birth control on thrombosis via the path going through the blood chemical. Birth control is a type-level positive cause of the blood chemical, which in turn is a cause of thrombosis. Does this show that birth control causes thrombosis via the blood chemical for the women in this population? Not necessarily. Suppose there are some women who have the blood chemical, though not as a result of having taken the pill. In these women, it is not the case that birth control causes thrombosis via producing the blood chemical, since their having the blood chemical does not depend on their taking birth control. So if we want to pick out the population of individuals for whom birth control influences thrombosis via this path, we should consider only those who had the blood chemical as a result of the pill and got thrombosis as a result of the chemical. In cases where these causal

[8] Eells changed his position between Eells, 1988 and Eells, 1991, and here I give the latter. In the former, he requires one to hold fixed causes of mediator *M* that are "simultaneous with or prior to the time of the cause" (99). One need not hold fixed causes of the mediator occurring after *C* but before *M*. In the book, however, Eells maintains that one must hold fixed all causes of the mediator (whenever they occur), provided they are not effects of *C*. Eells never notes his change in position. This is important, since Cartwright's primary objection is that Eells arbitrarily distinguishes between causes of the mediator that occur before and after *C* (89). By not acknowledging his change of position, Eells leaves the reader with the false impression that Cartwright was responding to the revised view.

[9] What follows is my best attempt at reconstructing the argument in Cartwright 1988.

relationships are irreducibly chancy, which women get the blood chemical as a result of the pill will not be determined by their properties alone. So to pick out the relevant set of individuals, we cannot appeal just to type-level facts about the women in the population (i.e. to their properties), but must also appeal to facts about singular causation.

Cartwright takes her argument not merely to show that we must appeal to singular causal facts in evaluating the effect of birth control on thrombosis via a path. She claims that even in evaluating the net effect of birth control on thrombosis, we must condition on mediators in the way she proposes. Since taking the pill token-influences the mediators differently in different women, averaging over mediators (by not conditioning on them where appropriate) obscures the "true causal role" (1988, p. 92) of birth control.

As Hitchcock (2001a) notes, the type/token distinction conflates (at least) two separate distinctions, both of which matter here. One distinction is that between actual causes and causal tendencies: Birth control may tend to cause thrombosis in a woman or population of women even if she or they never actually take birth control and get thrombosis as a result. Another distinction is between causal generalizations involving heterogeneous and homogenous populations. In a population where everyone is causally similar to Lisa, Lisa has the same tendency to get thrombosis as anyone else in the population. In contrast, if the effect of the pill on thrombosis differs among members of a heterogeneous population, then the tendency of the pill to cause thrombosis in Lisa may differ from that of other members of the population. The second distinction is crucial for Cartwright here. We must consider singular causal facts since without them we miss a form of within-population heterogeneity. But heterogeneity in what? Heterogeneity in which women actually have the blood chemical, or in which women have the tendency to get it?

Neither answer is satisfying. As Eells emphasizes, Cartwright's argument presupposes that all variation in which women get pregnant is irreducibly the result of chance. For each variable that causally influences pregnancy, Eells' theory requires it to be held fixed at the same value for every member of the population. For example, the population must be uniform with respect to whether the women are sexually active, as this influences their chance of pregnancy. With such factors fixed there is no further property determining which women get pregnant, and therefore no variation in their tendencies to get pregnant. So maybe Cartwright is referring to heterogeneity in which women actually get pregnant. But even if there is variation in which women get pregnant, this need not be variation in the *effect* of birth control on pregnancy, as there must be for her argument to succeed.

Arguably, the best way to respond to Cartwright would be to deny the existence of causal variation in populations with homogenous

background factors. This is not the route that Eells takes. Eells grants that there might be variation in the *token*-causal relationships that lead to women getting (or not getting) thrombosis via one path or the other (1991, p. 234-5), but notes that this variation is consistent with there being type-level homogeneity. Eells' most straightforward argument that type- and token-relationships may come apart is a variant of the phone call case. In 90% of all cases, your calling me causes the phone to ring, which in turn causes me to pick up the phone. In the other 10% of cases, there is a device that randomly causes the phone to ring (and for me to pick up the phone) whether or not you've called me. Now imagine a set of cases in which you call me and, coincidentally, every time you call the device activates and causes the phone to ring independently of your calling. Question: What is the effect of your calling me on my picking up the phone? Answer: calling me increases my probability of picking up the phone by .9. On Cartwright's approach, we would hold fixed the phone's ringing, since in every case it is token independent of your calling. This yields the result that your calling is not relevant to my picking up. Yet the requirement that we *must* hold fixed the mediator when it is token-independent of its cause throws out the baby with the bathwater. Type-level causal claims concern the relationship between variables in a representative sample. In such a sample, your calling me does cause me to pick up the phone 90% of the time. Yet Cartwright's account denies that there is any causal relevance relationship here.

While Eells' example shows why we should not condition on mediators in the way Cartwright suggests, it is controversial whether it shows that token causal facts come apart from type-level causal facts. Instead of maintaining that your calling is a type-level, but not a token-level, cause of the phone's ringing, one could argue that because the relationship between events is irreducibly probabilistic, your calling is neither a type- nor token-level cause of the ringing. Instead, it is a deterministic cause (of both varieties) of the *probability* of ringing (Hausman, 1998).[10] This strikes me as the best thing to say here, though I will not insist upon it. My point in revisiting the Eells/Cartwright debate is to illustrate how these philosophers came to link the analysis of multipath cases to issues of causal heterogeneity and the distinction between type and token causation. In the next section, I argue that these issues are not relevant to giving an account of path-specific effects. This account appeals not to variation across individuals, but to counterfactual values of mediators for an individual. I characterize individuals entirely in terms of their properties, so everything in the account can be cashed out in terms of causal tendencies.

[10] To make this view plausible, one must maintain that token causal claims are true in virtue of counterfactuals about the token relata.

3. Defining Path-Specific Effects

Let's now consider the thrombosis example using structural causal models. In figure 1, birth control (*X*) causes thrombosis (*Y*) both indirectly via pregnancy (*M*) and directly through mediators that were not included in the model. One could of course include mediators such as *blood chemical* in the model, but here I want to emphasize that one need not include a mediator along every causal path to evaluate the contributions of particular paths. I refer to *X* as the *treatment*, *M* as the *mediator*, and *Y* as the *outcome*. The causal path going through *M* is the indirect path and that consisting of a single arrow from *X* to *Y* is the direct path. Clearly, the notions of direct and indirect are model-relative.

The aim of *causal mediation* is to determine how much of the *total effect* – the effect going through all paths – depends on the contributions of particular paths. The contributions via the direct path are *direct effects* and those via the indirect path are *indirect effects*. The term "path-specific effects" encompasses both direct and indirect effects.

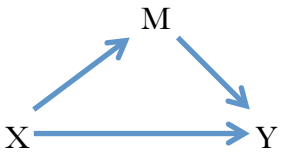

Figure 1 – *X*: Birth control
M: Pregnancy
Y: Thrombosis

The graph in figure 1 corresponds to two structural equations: that for the effect of the treatment on the mediator and that for the effect of the treatment and mediator on the outcome. The parameters in these equations can be identified via interventions or by conditioning on confounders. One might suppose that facts about path-specific effects could be straightforwardly read off from the structural equations. And in models with linear parameters, this is roughly the case. Suppose that the equations for the model in figure 1 were as follows:

$$(1) \quad M = aX + U_M$$
$$(2) \quad Y = bM + cX + U_Y$$

In such a case, the direct effect is *c* the indirect effect is given by *a* times *b* and the total effect by *ab* + *c* (Baron and Kenny, 1986).

The linear case obscures the true challenges. When the mediator and treatment do not contribute additively to the outcome, it is not clear it even makes sense to decompose the total effect into direct and indirect components. Consider the influence of the treatment on the outcome along the direct path. When the treatment and mediator interact, how the treatment influences the outcome along the direct path depends on the value of the

mediator, which itself depends on the treatment. So there is no neat way to divide the total effect into two additive components.

The challenge of providing an account of path-specific effects is not merely that of extending the definitions of the direct and indirect effect beyond the linear case. The core philosophical question is that of whether we can make literal sense of claims that an effect is transmitted via a path. Eells was willing to grant that the total effect of the pill on thrombosis would be different in a population of women for whom birth control has no effect on pregnancy. But he denies that the effect in such a population is relevant to understanding the effect of birth control on thrombosis in which there *are* two paths.

I now use Pearl's (2001) definitions for path-specific effects to address these challenges.[11] Let's begin with the direct effect. A glance at the model suggests an obvious way for finding the contribution of the direct path: one should intervene on the mediator. Doing so makes it so that the treatment cannot influence the mediator, and it will therefore influence the outcome only via the direct path. Yet defining a direct effect is not just a matter of saying that one must intervene on the mediator, since one must specify the value to which the mediator is set. Moreover, privileging any value or set of values of the mediator as being relevant to measuring the direct effect seems arbitrary. In intervening on the mediator, we are trying to determine what the effect would be of the treatment on the outcome if, contrary to fact, it had no influence on the mediator. But this question does not lend itself to a non-arbitrary answer; if the mediator no longer depends on its prior causes, what constraints are there on which values it might take on?

The question of how the treatment would directly influence the outcome independently of the value of the mediator cannot be answered in models with interaction. A question that can be answered is that of how a particular *change* in the value of the treatment would influence the outcome if it were only transmitted via the direct path. To see how focusing on changes helps, consider the variables in the model prior to changing the value of the treatment and prior to disrupting any of the paths. According to this model, the variables have the values they would have corresponding to the pre-change value of the treatment and the relevant structural equations. Now imagine what would happen were one to change the value of the treatment, and this change made no difference in what happened along the direct path. The question of what value the mediator would take on in this scenario can be unambiguously answered: it would have the exact same value (or distribution) that it would have were one not to have changed the value of

[11] Although I rely on Pearl's formulations, many of the key insights discussed here are already present in Robins and Greenland (1992).

the treatment. This is the key to defining the direct effect. Informally, it is the effect of changing the treatment from one value to another, while intervening on the mediator to make it behave as if there were no such change.

Let's make this more precise. Here Donald Rubin's (1974) potential outcome notation is useful. Potential outcomes are counterfactuals concerning how an individual would respond to a treatment. The potential outcome of receiving treatment $X=x$ on outcome $Y=y$ for individual i is denoted as follows:

$$(1)\ Y_x^i = y$$

(1) is a deterministic counterfactual saying that if i receives treatment level x, her outcome with respect to Y will be y (we will later generalize this to populations). For example, if $X=x$ is taking the pill and $Y=y$ is getting thrombosis, (1) says that were i to take the pill she would get thrombosis. The truth-value of this statement follows from the structural equation linking X and Y. In general, the properties of potential outcome expressions are derivable from structural causal models (Pearl, 2009, chapter 7). The utility of potential outcomes notation for explicating path-specific effects derives from the following two features. First, one can easily represent interventions on two (or more) variables as follows:

$$(2)\ Y_{x,m}^i = y$$

(2) says that $Y=y$ when one sets X to x and M to m via interventions. Second, one can nest potential outcomes:

$$(3)\ Y_{Mx}^i = y$$

(3) denotes that $Y=y$ when M is set to the value that it would take on were X to be set to x via intervention.

In potential outcomes notation, the total effect is given as follows:

$$(4)\ TE_{0,1}^i(Y) = Y_1^i - Y_0^i$$

This is the difference in the value of the outcome between the case where one receives the treatment and that where one receives the control. In evaluating the total effect, one does not, of course, intervene on the mediator. One could nevertheless use potential outcomes to indicate that the mediator is a function of the treatment:

$$(5)\ TE_{0,1}^i(Y) = Y_{1,M(1)}^i - Y_{0,M(0)}^i$$

In contrast, to evaluate the direct effect of going from X=0 to X=1, one does not want the mediator to respond to this change in the value of the treatment. One therefore holds it fixed to M_0 – the value it takes on in the control scenario:

$$(6) DE_{0,1}(Y) = Y^i_{1,M(0)} - Y^i_{0,M(0)}$$

This equation clarifies how although one disrupts the indirect path by intervening on the mediator, one does not try to make the system behave as if the treatment had *no* influence on the mediator. One intervenes on the mediator to make it behave as if it were still a function of the treatment, though one pretends that the treatment was maintained at its control value. By setting the mediator as a function of the treatment, one makes the subjects of the intervention behave as if they were governed by the same structural equations as the subjects in whom one evaluates the total effect. By setting the mediator to its control value, one makes the indirect path behave as it would in the absence of the change from X=0 to X=1.

It is possible to generalize the potential outcomes framework to cases involving populations of non-homogenous individuals. It is straightforward to represent the *probability* that Y would be y were X to equal x: $P(Y_x = y)$. Given a distribution of outcomes, one is often interested in the expected value of the outcome, $E(Y_x)$. While it is possible to define the total effect for an individual (as we have done), in practice one can only learn the *average* treatment effect (from here on out, I remove the superscript i to allow for average effects):

$$(7)\ ATE_{0,1}(Y) = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$$

The *direct* effect of going from X=0 to X=1 for a heterogeneous population is the expected change in the outcome when one changes the treatment from 0 to 1 while setting the mediator to M_0 *for each member of the population,* where individuals may differ in their values for M_0. In evaluating the direct effect, knowing the average value of the mediator is not sufficient. Populations with the same average value of the mediator may differ their direct effects.

What I have called the direct effect is sometimes called the *natural* direct effect, to indicate that one sets the value of the mediator to the value (or distribution) that it would take on when X=0. This may be contrasted with the *controlled* direct effect:

$$(8) CDE(m)_{0,1}(Y) = Y_{1,m} - Y_{0,m}$$

The controlled direct effect is evaluated by holding the mediator to some arbitrary value *m* for every individual in a population. It corresponds to what Hitchcock (2001b) calls the "direct effect". Yet, there are as many controlled direct effects as there are values of the mediator, and Hitchcock provides no guidance as to which ones are relevant in which contexts. All uses of 'direct effect' in what follows refer to the *natural* direct effect.

A consequence of the fact that path-specific effects are defined relative to *changes* in the value of the treatment is that the direct effect of going from X=0 to X=1 is different from that of going from X=1 to X=0. The key difference between these is that in the former one holds the mediator fixed to M_0 and in the latter one holds it fixed to M_1. Relative to $TE_{0,1}(Y)$, I refer to the former (eq. 6) as the *sufficient* direct effect and the latter as the *necessary* direct effect, which I define as follows:

$$(9) -DE_{1,0}(Y) = Y_{1,M(1)} - Y_{0,M(1)}$$

I have stipulated that the necessary direct effect corresponds to *negative* $DE_{1,0}(Y)$ so that with both the sufficient and necessary direct effects, one subtracts the potential outcome for the control scenario from the potential outcome for the treatment.[12]

As how we label the values of X is arbitrary, the distinction between necessary and sufficient may seem so as well. Nevertheless, once we define a total effect as going from one value to another, whether an effect is sufficient or necessary *relative* to that effect is not arbitrary. While $TE_{0,1}(Y)$ equals negative $TE_{1,0}(Y)$, and these are thus not independent quantities, stipulating a direction for the total effect allows one to treat one value of X as a default. When the default value of the treatment is X=0, we can call M_0 the default value of the mediator and $Y_{0,M(0)}$ the default value of the outcome. The sufficient direct effect is the amount by which the change in the treatment to its non-default value (X=1) would raise (or lower) the expected value of the outcome were all of the variables along the indirect path to remain at their default values. This is the sense in which the change in the value of X is only transmitted via the direct path. With the necessary direct effect, one starts with a system in which all the variables are at their *non*-default values and considers what would happen if one 'removed' the effect of the treatment (by changing it back to its default values) while maintaining the variables along the indirect path at their non-default values.

The counterfactuals given by the sufficient and necessary effects are relevant to different individuals. The sufficient direct effect would be relevant for considering a woman who does not take the pill, and wants to know what would have happened were she to have taken the pill, but it had no influence

[12] I am grateful to Malcolm Forster for pointing out the benefits of defining necessary effects in this manner.

on whether she got pregnant. The necessary direct effect is relevant to the woman who does take the pill and wonders what would have happened had she not taken the pill and doing so had not increased her chance of pregnancy.

I will now explain how to evaluate indirect effects. In cases where there is a mediator along every causal path, measuring indirect effects presents no problems beyond those already discussed with respect to the direct effect. By intervening on a mediator along every path aside from the path of interest, one can measure the indirect effect in the same way that one measures the direct effect. Novel difficulties arise in measuring the indirect effect for models without a direct path. There is no included mediator that one can intervene on in order to disable the direct path. It might seem impossible to measure the indirect effect without including such a mediator in the model and intervening on it. Yet, it is possible to do so.

As with direct effect, in evaluating the indirect effect one relies on a model's structural equations to determine the behaviors of variables under counterfactual circumstances. The structural equation linking the treatment to the mediator specifies what values the mediator would take on in the treatment and control scenarios. So to make the mediator behave as it would were it responding to a change in the treatment, we need not actually intervene to change the treatment, but can instead intervene on the mediator to behave as if it were responding to a change in the treatment. Accordingly, we can measure the indirect effect by holding the treatment fixed at its control value while varying the mediator to behave as it would in response to a change in treatment:

$$(10)\ IE_{0,1}(Y) = Y_{0,M(1)} - Y_{0,M(0)}$$

By holding the treatment fixed at X=0 one ensures that any unmeasured mediator along the direct path behaves exactly as it does when one does not receive the treatment. By comparing the values of the mediator for M_1 and M_0, one simulates how the mediator would respond to an intervention that changes the treatment from X=0 to X=1. While we distinguished between natural and controlled versions of the direct effect, there is no controlled version of the indirect effect, since there is no variable in the model to intervene on to identify the effect.

In the thrombosis case, the indirect effect is the effect that the pill would have on thrombosis via pregnancy, were it to have no influence through the direct path. Put differently, it is the effect that it would have if it influenced thrombosis only via pregnancy. In this case, we know that in addition to the pill influencing thrombosis via pregnancy, it also influences thrombosis via a blood chemical. Yet one does not need to know about the

blood chemical in order to measure the indirect effect.[13] When one sets the treatment to X=0, the (non-included) variable corresponding to the blood chemical takes on whatever value it does when one does not take the pill. The same is true for any other mediator along a path not going through *pregnancy*. Moreover, even if there were *no* mediators corresponding to the direct path, it would still be possible to identify the indirect effect. We tend to think that such cases of action-at-a-distance are impossible, or at least rare. It is nevertheless significant that our concept of the indirect effect does not entail the existence of omitted variables along the direct path.

We may distinguish between sufficient and necessary indirect effects, as follows:

Sufficient Indirect Effect: $IE_{0,1}(Y) = Y_{0,M(1)} - Y_{0,M(0)}$
Necessary Indirect Effect: $-IE_{1,0}(Y) = Y_{1,M(1)} - Y_{1,M(0)}$

The sufficient indirect effect compares the case in which taking the treatment only influences the outcome through the indirect path to the case where one does not take the treatment. The necessary indirect effect compares the case in which taking the treatment has no effect via the indirect path to the case where it makes a difference through both paths. The necessary indirect effect would be relevant to a woman who took the pill, did not get pregnant, and got thrombosis, and wants to know what her chance of thrombosis would have been had the pill not lowered her chance of getting pregnant.

Here I lack the space to discuss the range of cases in which path-specific effects are relevant. Here's one suggestive case (Pearl, 2001). In the US, being black may lead one to have fewer opportunities to develop the qualifications required for a certain job. But if being black causes one not to get a job *via* having fewer qualifications, this would not show that the hirer discriminated. The hirer discriminated if being black *directly* influenced one's not getting the job. Specifically, the relevant question is whether the applicant would have been more likely to get the job had she been non-black, but had the same qualifications she has as a result of being black. That is, is there a sufficient direct effect?

For reference, I provide a visual representation of the effects described here (figure 2). The arrows in the figures should be interpreted as in DAGs. The nodes have been replaced with empty circles for variables that take on their default values (X=0, M_0, $Y_{0,M(0)}$) and filled circles for variables with non-default values (X=1, M_1, $Y_{1,M(1)}$). Variables with other possible

[13] Although one need not know which mediating factors lie along the direct path, one must assume that none of them are causes of mediators along the indirect path. If any were, there would be an omitted common cause of the mediator and the outcome.

values are given a question mark. Each of the effects described corresponds to the result of subtracting the expected value of the outcome variable on the left from its expected value on the right. The nodes in the diagrams resemble those in David Lewis' neuron diagrams, although instead of the shading indicating the occurrence or non-occurrence of an event, the nodes correspond to whether a variable takes on its default value. Provided one bears this in mind, figure 2 serves as a convenient shorthand representation for the complex interventions required for measuring path-specific effects.

In cases without interaction, $TE_{0,1}$ is equal to the sum of $DE_{0,1}$ and $IE_{0,1}$. Yet the total effect is not, in general, the sum of the sufficient direct and sufficient indirect effects. To see this, consider the extreme case where either path would be sufficient for bringing about a certain value of the outcome. In such cases, adding together the sufficient direct and indirect effects would be double counting. Instead, the total effect is the sum of the sufficient direct effect and the *necessary* indirect effect:

$$(11)\ TE_{0,1} = DE_{0,1} + \left(-IE_{1,0}\right) = DE_{0,1} - IE_{1,0}$$

To see what's going on here, imagine that the sufficient direct effect is less than the total effect. Since the direct path is unable to account for the full effect, the indirect path must 'pick up the slack' so to speak and the necessary indirect effect indicates the amount that it must contribute in order to get to the full effect. Alternatively, the total effect may be decomposed into the sum of the sufficient indirect effect and the necessary direct effect:

$$(12)\ TE_{0,1} = IE_{0,1} - DE_{1,0}$$

The validity of (11) and (12) may be easily verified by consulting the relevant diagrams in figure 2.

In cases without interaction, $-IE_{1,0}$ is equal to $IE_{0,1}$. Substituting this into (11) yields:

$$(13) TE_{0,1} = DE_{0,1} + IE_{0,1}$$

While the decomposition in the additive case is more easily grasped than those in (11) and (12) it is a special case. Unless one can rule out interactions among variables, one must use the more complicated decompositions presented here.

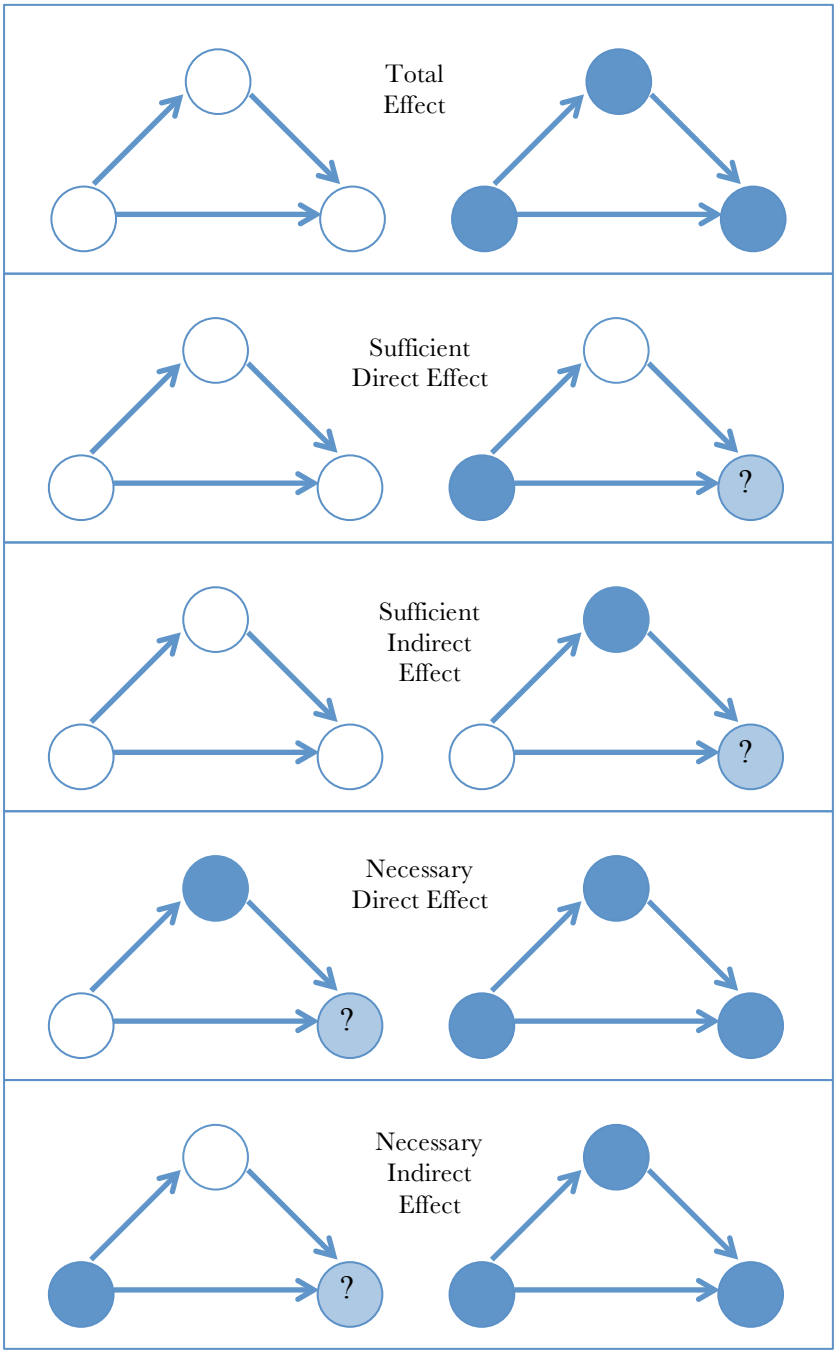

Figure 2: White circles indicate X=0, M_0, or $Y_{0,M(0)}$. Shaded circles indicate X=1, M_1, or $Y_{1,M(1)}$. Outcomes with values that need not match those mentioned have question marks. Each effect is derived by subtracting the (expected) outcome on the left from that on the right.

In general, describing path-specific effects in terms of the decomposition of the total effect into direct and indirect contributions is misleading. While one can give a non-additive decomposition, this does not change the fact that one cannot fully separate the contributions of the paths in cases with interaction. A better way to think about path-specific effects is as representing the way that a path would transmit a particular change in the treatment were this change not transmitted via the others. These two ways of thinking about path-specific effects do not come apart in additive models, but diverge in models with interaction. It is the relativity of path-specific effects to *changes* in the treatment that determines the counterfactual scenario relevant for evaluating them – it is the scenario in which the behaviors along the other paths remain unchanged.

In focusing on changes in the value of the treatment, one introduces an asymmetry that is not present in the structural equations. One treats the variable values corresponding to one value of the treatment as being defaults, and variable values corresponding to another as changes from the default.[14] Here I take no stand on whether the treatment of certain values as defaults is a pragmatic choice, or whether it reflects something deeper about causation (Maudlin, 2004). The reason why talk of defaults helps in understanding mediation is that it highlights why path-specific effects corresponding to a particular change in the treatment differ from those corresponding to the reverse change.

The relativization of path-specific effects to changes in the treatment goes hand in hand with the evaluation of these effects in populations where the structural equations are "preserved". While intervening on the mediator destroys the treatment-mediator relationship, I have explained how it is possible to do so to simulate the case where the structural equations are intact and only the information the paths transmit regarding the value of the treatment differs. In the next section, I argue that it was this feature of path-specific effects that was beyond the reach of probabilistic theorists.

4. Path-Specific Effects and Probabilistic Causality

For path-specific effects measured in one set of individuals to correspond to a quantity that can be said to play a role in bringing about the total effect in a distinct set of individuals, one needs a basis for saying that these two distinct sets of individuals may be used to evaluate the same total effect. Our discussion has revealed such a basis. The individuals in whom one measures

[14] A reviewer suggests that the definitions provided could be made more concrete by applying them to a linear model with interaction. Although I lack space to do so here, I direct interested readers to Pearl (2014), section 3.3. There one can verify that $DE_{0,1}$ depends on b_0, which is the baseline value of the mediator when X=0. This illustrates the relativity of path-specific effects to default states of off-path variables.

the path-specific effects are made to behave in such a way as if they were still governed by the same structural equations that account for the total effect. It is unsurprising that probabilistic theorists missed this similarity. They distinguished between groups of individuals by referring to sets of background factors, and the populations in which one measures total and path-specific effects differ in their background factors.

Probabilistic theorists' emphasis on background factors as opposed to structural equations helps explain why they thought that separating the contributions of paths might require an account of token causation. For them, to explain the total effect in some individuals by referring to individuals with different background factors would be to change the topic. Yet amongst individuals with the same background factors there is, of course, no variation in any causally relevant properties. So (it seemed) any further distinction regarding how a cause influences its effect in one way rather than another must appeal to token-level causal facts that do not obtain in virtue of the properties of individuals.

Structural equations allow us to model path-specific effects in terms of how individuals would vary under relevant counterfactual circumstances rather than in terms of some alleged causal variation among individuals. This is why, more generally, the question of whether an effect averages over heterogeneous populations is orthogonal to whether it is a genuine effect. To the extent that we can measure average effects in heterogeneous populations, it is because this effect averages over the total effects for the individuals in that population (Weinberger, 2015). What matters are an individual's outcomes for counterfactual values of the treatment, rather than whether there is actual variation among individuals. A similar point holds for path-specific effects. These effects, by design, average over the relevant values of M_0 and M_1 for individuals in a population. This feature of path-specific effects leads to practical complications (Imai et al., 2011). But this does not undermine the conceptual point that the definitions for path-specific effects, like those for total effects, need not distinguish between homogenous and heterogeneous populations.

In short, when it comes to understanding multipath cases, theorists of probabilistic causality were on the wrong track. The definitions provided for the four path-specific effects do not rely on the distinction between type and token causation, however it is understood. The definitions apply to both homogenous or heterogeneous populations, and the total effect is not an average over different path-specific effects. The definitions for path-specific effects do not entail that any variable values are instantiated, and the quantities they require may be read off from the structural equations. Accordingly, they reflect causal tendencies rather than facts about actual causation. Whatever motivations there are for distinguishing between type

and token causation, providing an analysis of path-specific effects is not one of them.

Accounts proposed since the heyday of probabilistic theories have been no better at clarifying the relationship between the total effect and path-specific effects. Consider Hitchcock's (2001b) distinction between net and component effects. His notion of a component effect is very similar to that of a path-specific effect – in some models it corresponds to the controlled direct effect. Yet, he does not aim to show how component effects *relate* to net effects. His point is precisely that component effects and net effects are two distinct concepts and our discomfort in saying whether birth control prevents thrombosis results from not disambiguating them. Birth control may be a net preventer of thrombosis while also promoting thrombosis via the blood chemical. Although there is a question for Hitchcock as to why one cares sometimes about net effects and at other times about component effects, the question of how the former relate to the latter does not arise. In contrast, I have defined path-specific effects in relation to the total effect.

5. Conclusion

Although Pearl's definitions for direct and indirect effects are now several years old, the concept of a path-specific effect explicated here substantially diverges from that of prior philosophical accounts. This is evident from the fact that the present analysis allows for quantitative definitions of the contributions of paths, even in cases with interaction. Yet the philosophical importance of path-specific effects derives less from their enabling us to use numbers than from their allowing us to clarify what these numbers measure. We no longer need to treat path-specific effects as unrelated to the total effect, but can specify their counterfactual contribution to it.

Given that I have argued that probabilistic theories were on the wrong track in thinking about multipath cases, why discuss them at all? The reason that the probabilistic theorist's misdiagnosis of the Hesslow case is interesting is that in many respects these theories resemble those relying on structural causal models. In linking causation to probability, the recent literature fulfills many of the ambitions of the probabilistic project. Yet structural equations theories reorient us away from thinking about causation in terms of correlations in populations and towards thinking about it in terms of counterfactual relationships among variables. Just as in physics, new theories distinguish themselves from their predecessors by rendering the anomalous lawful, so too new conceptual schemas are distinguished by their rendering the unintelligible sensible. By introducing structural equations, we transform the transmission of influence along a path from something that must be explained away into something explicable using the same resources that suffice for understanding the total effect: the structural relationships among variables.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173–1182.

Cartwright, N. (1979), "Causal Laws and Effective Strategies", Nous 13: 419-37.

Cartwright, N. (1988). Regular associations and singular causes. In *Causation, Chance and Credence* (pp. 79-97). Springer Netherlands.

Cartwright, N. (1989). Nature's Capacities and their Measurement.

Dupre, J. (1984). Probabilistic causality emancipated. *Midwest Studies in Philosophy*, *9*(1), 169-175.

Eells, E. (1987). Probabilistic causality: reply to John Dupré. *Philosophy of Science*, 105-114. ISO 690

Eells, Ellery, (1991). *Probabilistic Causality*, Cambridge, U.K.: Cambridge University Press.

Eells, E. and Sober, E. (1983), "Probabilistic Causality and the Question of Transitivity", Philosophy of Science 50: 35-57.

Fitelson, B. and Hitchcock, C. (2011): 'Probabilistic Measures of Causal Strength', in P. McKay Illari, F. Russo and J. Williamson (eds), Causality in the Sciences, Oxford: Oxford University Press, pp. 600–27.

Good, I. J. (1961-2), "A Causal Calculus I-II", British Journal for the Philosophy of Science 44: 305-18; 45: 43-51. Errata and Corrigenda, 49: 88.

Hausman, D. M. (2010). *Probabilistic causality and causal generalizations* (pp. 47-63). Springer Netherlands.

Hesslow, G. (1976), "Discussion: Two Notes on the Probabilistic Approach to Causality", Philosophy of Science 43: 290-292.

Hitchcock, C. R. (1995). The mishap at Reichenbach fall: Singular vs. general causation. *Philosophical studies*, *78*(3), 257-291.

Hitchcock, C. (2001a). Causal generalizations and good advice. *The Monist*, 218-241.

Hitchcock, C. (2001b). A tale of two effects. *Philosophical Review*, 361-396.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765-789.

Maudlin, T. (2007). *The metaphysics within physics*. Oxford University Press.

Otte, R. (1985), "Probabilistic Causality and Simpson's Paradox", Philosophy of Science 52: 110-125.

Pearl, Judea, (2001). "Direct and Indirect Effects," *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, pp. 411–420

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge:
Cambridge University Press.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143-155.

Rosen, D. A. (1978). In defense of a probabilistic theory of causality. *Philosophy of Science*, *45*(4), 604-613.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Skyrms, B. (1980), Causal Necessity. New Haven and London: Yale University Press.

Sober, E. (1984). Two concepts of cause. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (pp. 405-424). Philosophy of Science Association.

Spirtes, Peter, Clark Glymour, and Richard Scheines, (2000). *Causation, Prediction and Search*, second edition, Cambridge, MA: M.I.T. Press.

Suppes, P. (1970), A Probabilistic Theory of Causality. Amsterdam: North Holland Publishing Company.

Weinberger, N. (2015) If Intelligence is a Cause, It is a Within-Subjects Cause. *Theory and Psychology 25* (3): 346-361

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press